引用格式：杨小康，许岩岩，陈露，等. AI for Science：智能化科学设施变革基础研究. 中国科学院院刊, 2024, 39(1): 59-69, doi: 10.16418/j.issn.1000-3045.20230820004.
Yang X K, Xu Y Y, Chen L, et al. AI for Science: AI enabled scientific facility transforms fundamental research. Bulletin of Chinese Academy of Sciences, 2024, 39(1): 59-69, doi: 10.16418/j.issn.1000-3045.20230820004. (in Chinese)

AI for Science：智能化科学设施变革基础研究

杨小康[1] 许岩岩[1] 陈 露[1,2] 王韫博[1] 高 岳[1] 田济东[1] 俞 凯[1,2] 金耀辉[1] 梅 宏[3*]

1 上海交通大学 人工智能研究院 上海 200240
2 上海交通大学 计算机科学与工程系 上海 200240
3 北京大学 信息科学技术学院 北京 100871

摘要 近年来，人工智能（AI）在前沿科技领域取得了诸如AlphaFold2、核聚变智能控制、新冠药物设计等诸多令人瞩目成果，表明AI for Science正在成为一种新的研究范式。实现智能时代的基础科学源头创新及其下游重大技术创新，需破解2个方面的核心问题：（1）如何利用新一代AI（特别是生成式AI及大模型）的通用性和创造性推动新范式的形成；（2）如何利用AI实现对传统科学设施的赋能与改造。文章提出一种智能化科学设施的建设构想，兼顾“高度智能化的科学新设施”和“AI赋能已有科学大设施”2个层面的需求，构筑AI for Science的科学设施体系，形成科学领域大模型、生成式模拟与反演、自主智能无人实验及大规模可信科研协作等创新功能，加速重大科学发现、变革性物质合成，以及重大工程技术应用。

关键词 人工智能驱动科学研究，生成式人工智能，语言大模型，科学设施

DOI 10.16418/j.issn.1000-3045.20230820004

CSTR 32128.14.CASbulletin.20230820004

1 科学研究新范式：AI for Science（科学智能）

1.1 科学研究范式的演进

人工智能（AI）在前沿科学与技术领域的应用已经取得了令人瞩目的重大成果。在生物领域，2021年《科学》杂志将AlphaFold2评选为“2021年度十大科学突破”榜首[1]；在物质领域，AI实现了核聚变托卡马克装置的等离子流高效控制[2]；在药物领域，AI加速了新冠药物设计[3]。众多的科学突破和国内外的发

*通信作者
资助项目：上海市人工智能重大专项（2021SHZDZX0102）
修改稿收到日期：2023年12月26日

展趋势表明，AI for Science 正在成为一种新的研究范式。

2007年，图灵奖得主Jim Gary曾经用“4种范式”描述了科学发现的历史演变，即实验观察、理论推导、模拟仿真、数据驱动（即数据密集型科学发现）（图1）。几千年前，人类通过实验观察和经验来描述自然现象，形成经验范式（第一范式）。几百年前，科学家开始使用理论推导进行科学研究，基于理论和模型解释自然现象，如17世纪牛顿定律的诞生，这是理论范式（第二范式）。20世纪中叶以来，科学家通过计算机模拟仿真复杂现象，进行理论验证，形成了计算范式（第三范式）。近10多年来，人类进入大数据时代，通过大数据分析去理解、逼近复杂系统，得出之前未知的科学理论，形成了数据密集型范式（第四范式）。然而，由于受限于数据采集与模拟空间，即便在数据密集型的科学研究范式下，科学假设依旧由科学家的专家经验主导；同时由于缺乏有效的数据开放机制和实验的局域性，制约了大规模、跨学科科研活动的开展。

最近若干年，深度学习技术，特别是生成式AI的迅猛发展，使得学术界可以利用深度学习建模和挖掘高维科研数据，捕捉多模态数据背后的科学规律，同时借助数据生成的方式，突破实验观测数据的有限性与数值模拟的理论限制，拓展科学假设的空间。多模态语言大模型潜在的文献理解总结、实验方案生成等全新能力，结合无人实验系统与科学数据开放平台，可促使科学研究迈向以“平台协作”为主要特征的新模式。微软剑桥研究院院长Chris Bishop等将AI for Science称为驱动科学研究的第五范式（利用AI和机器猜想来进行科学发现的新方法）[①]。与前4种范式不同，它不仅依赖于已有数据和方程，而且能够通过机器学习模拟自然现象，推断出某些未知的规律，提高科学研究的效率和准确性，探索更广阔的可能性空间，其中包括问题空间和解空间。

1.2 国内外对AI for Science的激励措施与计划

鉴于AI可能为科学发展带来的巨大助力，英国提出“用于科学研究和政府管理的人工智能”（AI for Science and Government）计划，旨在应用AI解决现实自然/社会科学问题。美国国家科学基金会（NSF）于2021年开展了一项AI4Science研究计划，使用AI推动数学和物理学领域科学发现。2022年，法国国家科学研究中心（CNRS）成立“人工智能与科学研究双向驱动的跨学科中心”（The Artificial Intelligence for Science, Science for Artificial Intelligence Center, AISSAI），推动不同领域间的交流与合作，拓展AI在科学研究中的应用。从2019年起，中国科学院及其所属研究院所在多个学科领域开展了AI for Science相关研究，涵盖生物医学、材料科学、计算物理及量子计



图1　科学研究范式的转变

Figure 1　Paradigm shift in scientific research

① Chris Bishop. 科学智能（AI4Science）赋能科学发现的第五范式. (2022-07-07). https://www.msra.cn/zh-cn/news/features/ai4science.

算等多个领域。

2017年，国务院在印发的《新一代人工智能发展规划》中指出："聚焦人工智能重大科学前沿问题，兼顾当前需求与长远发展，以突破人工智能应用基础理论瓶颈为重点，超前布局可能引发人工智能范式变革的基础研究，促进学科交叉融合，为人工智能持续发展与深度应用提供强大科学储备"。2023年3月，科学技术部会同国家自然科学基金委启动"人工智能驱动的科学研究"专项部署工作，推进面向重大科学问题的AI模型和算法创新，发展针对典型科研领域的AI for Science专用平台，布局AI for Science研发体系，逐步构建以AI支撑基础和前沿科学研究的新模式，加速我国科研范式变革和能力提升。

2 智能化科学设施

虽然AI在特定科学领域的应用已经取得了令人瞩目的重大成果，如AlphaFold2[1]、核聚变智能控制[2]、基于AI规划的全自动有机合成平台[4]、分子动力学模拟[6]及新冠药物设计[3]、生成式材料反向设计[7]等，但是这些工作仍然处于特定研究组对特定问题的科研模式，学科的跨度、研究场景的规模、研究方案及其结果的可复现性均受到限制，"平台协作式"的AI for Science科研模式及设施体系尚有待建立。

*实现智能时代的基础科学源头创新及其下游重大技术创新，急需破解2个核心问题。① 如何面向AI for Science的研究范式，建立全新的科学智能大设施？*从而系统性、整体性地释放新一代AI（特别是生成式AI及大模型）在基础科学领域的创造性和通用性，实现自发假设生成、自动规律推演、自主无人实验、自驱可信协作等创新功能，推动超大规模、高速迭代的科学探索？*② 如何利用新一代AI实现对传统科学设施的赋能？*科学研究活动主要包括由科学家提出问题和假设、由实验人员进行检验和验证、通过科研机构与出版商进行科研成果和数据传播与共享等环节。在传统的科学设施和研究范式下，整个科研流程存在科学问题沟通难、科学实验操作难、科学数据共享难等困难（图2）。一些高精尖的大型科学设施及其科研环境高度复杂，此类困难尤为突出。利用新一代AI实现"科学问题（科学家）—实验设备（实验员）—科研数据及文献（科研机构及中介）"高效闭环，不仅是新建的科学设施需要具备的标配，更是在已有的科学设施升级改造过程中的新需求和新机遇。

针对上述问题，2023年4月，上海交通大学人工智能研究院团队在浦江创新论坛"AI for Science专题论坛"上提出了建设"智能化科学设施"（AI enabled Scientific Facility，AISF）的构想。智能化科学设施的总体思路是瞄准世界科技前沿与国家重大需求，兼顾"创建高度智能化的科学新设施"和"赋能已有科学大设施"2个层面，构筑AI for Science的科学设施体系，加速重大科学发现、变革性物质合成及重大工程技术应用（图2）。

智能化科学设施综合运用生成式AI、语言大模型、大数据、区块链等前沿技术，形成人在环路的科学智能大设施3层体系架构（图3）。*① 基础支撑层*，通过高性能计算、算力网，形成算力支撑；*② 科学模型层*，构建跨学科、跨模态的科学大模型，以及"AI科研助手"；*③ 实验应用层*，通过AI操作机器人、智能实验环境，实现自主无人实验和多方科研协作。

在3层架构基础上，智能化科学设施可形成传统范式所不具备的4个主要新功能（图4）：*① 科学大模型*，实现跨模态的科研内容生成、文献综述生成、科学任务自动拆解及实验方案自动生成等能力，进而构造具有较高综合科学能力的"AI科研助手"系统模型；*② 生成式模拟与反演*，提供复杂流体、多物理场、复杂物质结构等科学现象生成及其AI加速的超大规模模拟能力，缓解维度灾难（curse of dimensionality），激发科学直觉；*③ 高通量自主无人实验*，将自动化实验室和AI模型结合，实现合成化

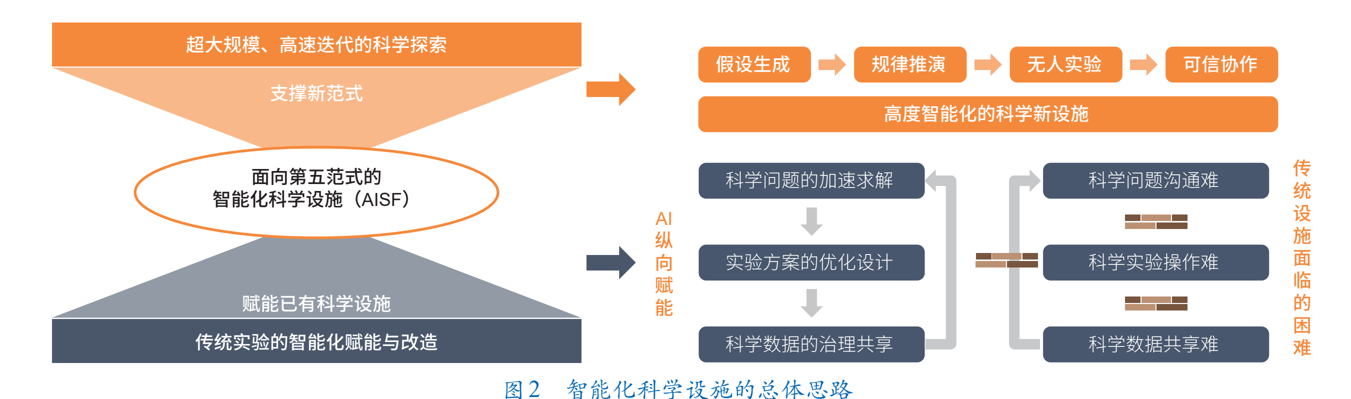


图2　智能化科学设施的总体思路

Figure 2　Overall approach of AI enabled scientific infrastructure

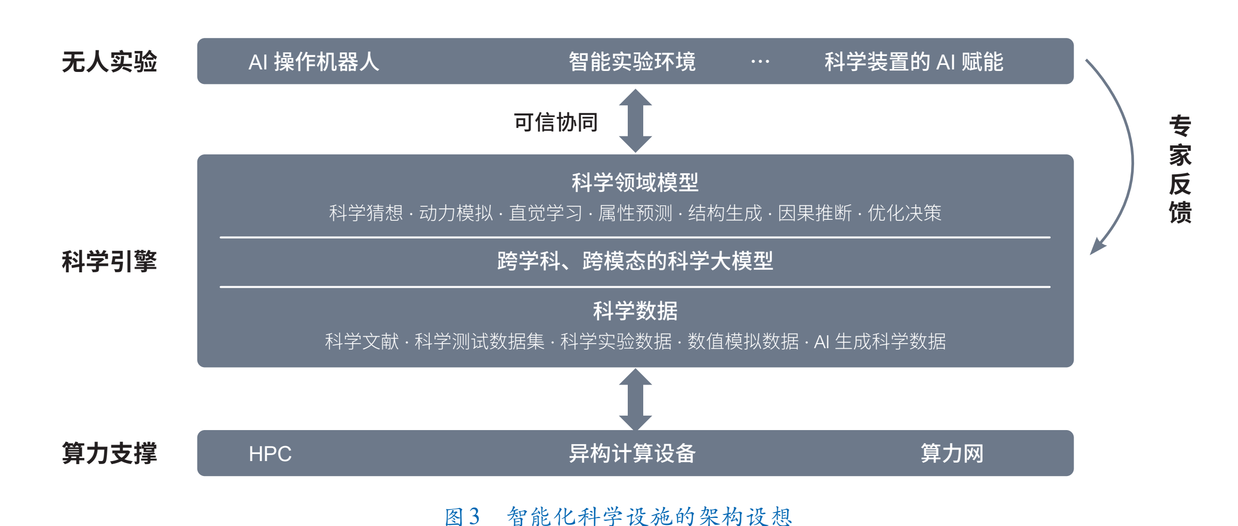


图3　智能化科学设施的架构设想

Figure 3　Conceptual architecture of AI enabled scientific infrastructure



图4　智能化科学设施的创新功能

Figure 4　Innovative functionalities of AI enabled scientific infrastructure

学、合成药物、材料基因组等领域的“干湿闭环”自主实验验证；④ 大规模可信科研协作：通过区块链、群体智能等技术，实现科学模型及数据集的链上溯源、确权、共享、流通，加速科学新思想和新方法的涌现。

在上述架构和功能基础上，以人类科学家和科学问题为中心，构筑“AI科研助手—AI操作机器人—智能实验环境—可信多方协作”的人机物协同科研空间，横向支撑超大规模的“假设生成—规律推演—无人实验”高速迭代、推进高度智能化的 AI for Science 科研范式转变，纵向优化基础科学大设施、赋能传统“实验观察—理论建模—数据分析”科研流程。

3 科学大模型

以 ChatGPT 为代表的对话式大语言模型被视为颠覆性的新一代 AI 技术。在基础大语言模型之上，通过科学大数据训练科学知识与能力，通过“科学家在环

路”的强化学习实现科学规范伦理对齐，打造适用于科学研究的大模型，有望形成AI与基础科学之间交互强化、回旋加速的创新机制与方法体系。

（1）AI正向驱动基础科学。一方面，利用AI提高科学研究的速度和准确性，以及对科学知识的融汇能力，探索更广阔的科学假设空间，促进多学科深度交叉和重大科学发现。另一方面，AI成为传统科研范式的有益补充，可有效提升实验观察、理论推导、仿真模拟、数据驱动等能力。

（2）基础科学反向驱动AI。① 科学大模型将为领域大模型的发展树立技术标杆。在主要面向人类自然语言和常用编程语言的大语言模型基础上，科学AI大模型还要面向数学公式、物理方程、化学分子式、材料结构、基因序列等跨学科、跨模态的科学语言和文献，可以对其他大模型的发展起到示范作用。② 相较于面向互联网应用的基础大模型，面向特定领域的科学大模型拥有更快的训练和迭代速度。同时，科学大模型的服务对象为科研工作者和高校学生，其使用过程中将产生海量的科学问题及回答，从而积聚科学知识、引导大模型生成新的科学假设。③ 当前大模型备受争议的是大数据、大算力导致的高能耗。这需要从开源、节流2个方面去解决。从开源角度来看，需推动能源科学技术的进步，进一步提升人类汇集、产生、存储、输送、使用能源的效率，这也是AI可以发挥重要作用的领域。从节流角度来看，需努力降低计算系统的能耗，追求绿色节能。随着计算所需的能源成本降低、AI计算体系架构的逐步完善，无论是基础大模型，还是科学大模型，都将进一步释放出更大的威力。

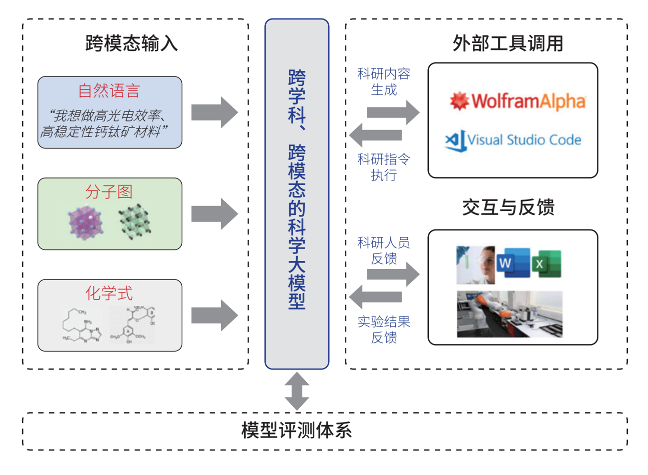


图5　科学大模型整体架构图

Figure 5　Architecture diagram of scientific large-scale models

为构建科学大模型（图5），我们需要在基础大语言模型之上形成4种专门能力，并构建对应的评测基准。① 跨学科跨模态统一输入的能力。语言大模型逐渐被应用到生物医学[7]、材料学和化学[8]等多个学科领域，已经可以较好地实现实体识别、关系抽取、领域分类等文本处理任务。在科学研究产生的数据中，除了文本数据，一般还有公式、图表、分子式等多模态数据，如何实现这些跨模态数据的统一输入，进而对跨学科的专业知识进行联合建模，是建立科学大模型亟待解决的挑战问题。② 有效调用外部科学工具的能力。大模型虽然拥有较强的语言理解、语言生成及复杂推理能力，但仍然存在输出看似合理、实则错误内容的问题。相比基础大语言模型，科学大模型对生成内容的专业性、准确性都提出了更高要求。解决此类问题的一种可行技术途径是将大模型作为规划与推理引擎，通过调用不同类型的外部科学工具，以提高可信度和准确性。③ 模型持续反馈和进化的能力。人类科学家可以依靠自身的经验积累和外部的持续反馈，不断提升科研能力。类似地，科学大模型一方面可以利用科研人员的高质量反馈，完善科学领域专业知识，增强科学知识建模与推断能力，提高科学内容生成、文献归纳推理等方面的水平；另一方面，可以利用实验反馈来不断提升其科学假设生成和实验方案优化的能力。④ 幻觉（hallucination）消除能力。目前，大模型仍然存在幻觉现象（即模型生成的内容与现实世界事实或用户输入不一致的现象），难以满足

科学研究的精度要求。如何有效筛选专家知识，并利用高精准度的专业知识进行强化学习，实现大模型的迭代和更新，是消除科学幻觉的重要方法。此外，基于基座模型的定向优化和集成，在特定领域提升模型的可信度和精准度，也是修正科学幻觉的重要途径。⑤ 科学大模型的评测基准。为了对科学大模型的能力进行准确评估，并促进其快速迭代，亟需在科学知识和常识的基础上建立覆盖学科面较广的科学知识测试基准，一方面测试科学大模型的跨模态、跨学科数据理解和建模能力；另一方面评测科学大模型是否具有利用科学工具准确鲁棒地完成复杂任务的能力，以及是否具有较强的拒绝生成虚假和有害内容的能力。

综上所述，科学大模型作为"AI助手"辅助人类科学家进行科学研究，需要具备跨学科的知识背景、处理跨模态数据输入、掌握外部科学工具的调用能力，结合反馈与评测，实现持续进化。另外需要强调的是，科学大模型作为人类的"AI科研助手"提出"科学假设"，仅仅是给人类科学家的建议，需要认真甄别后方可进行后续的科学论证或实验探索。

4 生成式模拟与反演

用仿真及计算机模拟工具实现"理论—现象"之间的推演是科学研究的重要范式之一。仿真空间从人类感官的直觉上建立了假设空间和观测空间的桥梁（图6）。在核物理等规律及理论明确、数据质量高、观测成本高的应用场景中，计算机仿真越来越多地作为实验的有效补充。然而，传统的基于数值计算的模拟方法在很多现实场景中至少存在求解速度与精度2个方面的局限：① 迭代计算速度慢，特别是大规模的科学问题，对算力需求巨大；② 诸多复杂科学现象的底层理论模型尚不完备，只能采取近似或忽略复杂高阶物理关系的方式进行模拟，使得仿真可能产生有悖于实际观测的结果。

生成式AI技术有望克服求解速度和精度上的局限性。① 以生成式神经网络为基础结构，将对复杂系统的数值求解问题转化为数据拟合问题，建立起从假设空间到仿真空间的高效映射，从而加速求解。② 利用生成式渲染技术，实现从仿真空间到观测空间的科学现象表观生成，进而实现横跨3个空间的闭环学习，驱动规律反演（图6）。

4.1 从假设空间到仿真空间的科学规律映射

对于理论模型相对完备的科学现象，生成式AI求解方法可以形成假设空间到仿真空间的映射，加速方程求解。关键技术难题在于如何将理论模型尽可能准确、有效地嵌入到机器学习求解器的神经网络训练过程中。一种方法是用传统数值模拟器产生的仿真数据训练神经网络，间接实现科学规律嵌入[9]。另一种更为直接的方式是将数学方程转化为神经网络的损失函数，从而利用科学先验使得机器学习求解过程快速收

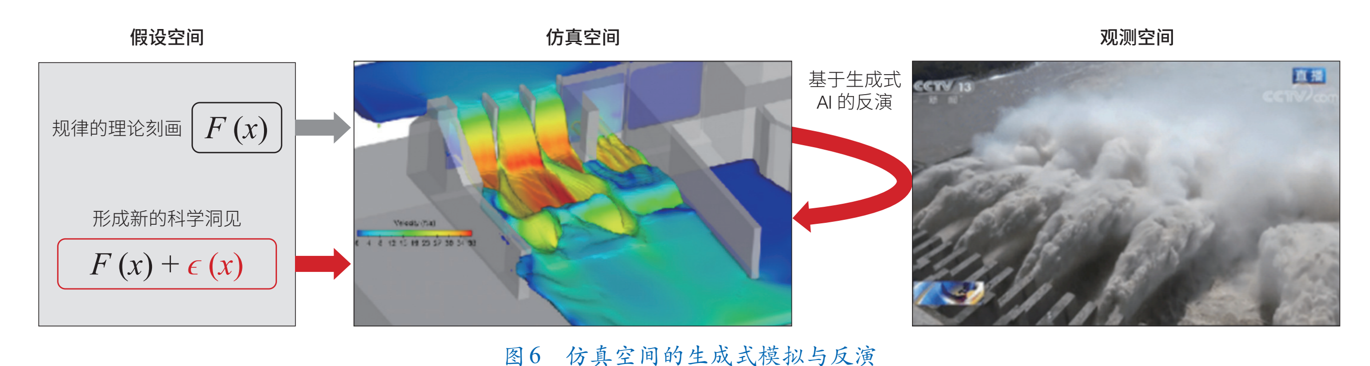


图6　仿真空间的生成式模拟与反演

Figure 6　Generative AI based simulation and inversion

敛到理论解附近[10]。然而，上述生成式AI求解方法在训练数据层面严重依赖数值模拟器的精度，对理论模型尚不完备的复杂科学现象，往往存在训练数据与真实数据存在偏差，边界条件与方程假设过度化简等系统性差异，在此基础上训练的机器学习求解器，不可避免地存在误差累积的问题。

4.2 “假设—仿真—观测”三空间的闭环学习

针对理论模型尚有待完善的复杂科学现象，生成式AI技术还可以被用来学习仿真空间到观测空间的数据映射关系[11]，有能力生成在统计学意义上符合观测分布、在感官上更逼近现实的信号与信息，从而允许科学家根据实际观测数据反演仿真空间中的状态“后验信息”，通过与仿真得到的“先验信息”进行比对，对已有的科学规律进行修正，甚至通过拓展假设空间，发现新的科学现象。

目前，生成式AI技术已被广泛验证可以极大地加速超大规模科学计算问题的求解，缓解维度灾难。以流体模拟为例（图6），“神经流体”（NeuroFluid）[12]提出的生成式AI技术的反演方法，以机器学习求解器的流体粒子仿真结果驱动基于“神经辐射场”（NeRF）的图形渲染，将数据中隐藏的物理规律从观测空间映射回容易被科学家理解分析、可以用状态参数调控的仿真空间，实现从自然图像中高精度地推理复杂几何结构周围的流体运动。但生成式AI技术的可解释性与鲁棒性还缺乏足够的理论保障。未来需要重点突破复杂流体、多物理场、复杂物质结构等科学规律嵌入到机器学习求解器的问题，以及探索如何进一步利用生成式AI技术进行“假设—仿真—观测”3个空间之间的推理，尤其是针对理论模型仍不完备的科学问题，激发科学直觉，实现理论模型的完善或者修正。

5 自主智能无人实验系统

自主智能无人实验系统旨在将AI和机器人领域的技术成果与科学实验进行结合，通过无人化、标准化、大规模的实验流程，提升实验效率和可复用性。美国麻省理工学院（MIT）的Coley等[4]于2018年提出AI规划的自动化学合成工作流，中国科学技术大学江俊团队[5]于2023年提出基于文献阅读的机器人自动合成装置。在此基础之上，智能化科学设施强调形成“开放协作、人在环中”的人—机融合实验模式（图7）。其中，无人实验操作平台及智能化系统的任务流程主要包括3个步骤：① 方案自动优化，在所提出科学猜想的基础上，自主进行模型设计和方案优化；② 自主任务规划，整合科学大模型，将实验和反应流程转化为形式化的机器人操作指令，完成实验操作的全流程自主规划；③ 无人实验操作，基于机器人的无人实验操作平台执行操作指令，开展自主无人实验，研究人员可通过人机交互接口远程监控实验状态。在无人实验操作平台及智能化系统的基础上，扩展形成安全、并行、协作的大规模开放实验平台。

5.1 无人实验的操作平台

根据操作精度的不同，无人实验操作可分为微观和宏观2种空间尺度。微观尺度下的无人实验的操作对象通常为活细胞、蛋白质等微小颗粒，多采用固定实验平台，其核心问题为提升高精度操作通量。宏观尺度的无人实验操作侧重实验流程的完备性，以搭载机械臂的可移动机器人为主在实验设备之间自主移

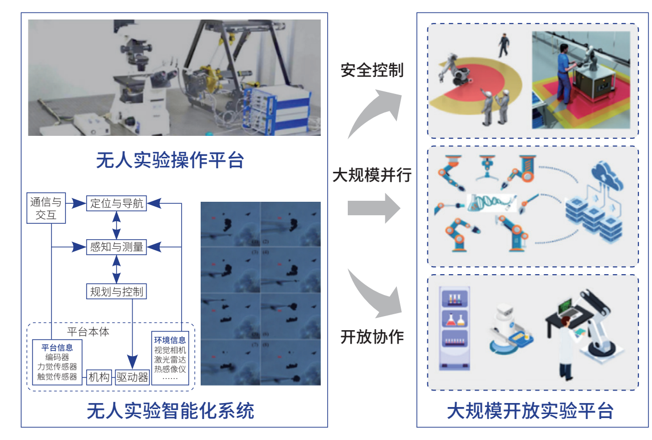


图7 自主智能无人实验系统

Figure 7 Autonomous unmanned experiment system

动，完成多任务全流程的自动化实验。此外，人形机器人能够在人—机融合协作环境中以更直接的方式实现高精度灵巧操作技能的模仿学习，有望进一步提升无人实验操作平台机器人的灵活性和自主性。

5.2 无人实验的智能化系统

在硬件本体平台的基础上，构建智能化系统软件是自主智能无人实验平台的另一个基础问题，通过软件驱动自身状态感知、外部环境感知、移动导航、仪器定位、实验操作规划和控制执行等流程。深度强化学习与模仿学习可以通过环境交互的经验轨迹或专家示教轨迹进行自主学习，构建观测信息与最优动作之间的映射关系[13]。近年来，ChatGPT等语言大模型为无人实验智能化系统的设计带来新的技术途径，一方面可以将语言指令和外界观测序列以端到端的方式映射为机器人动作指令；另一方面可以接收并理解人类输入的语言指令，自动将任务分解为若干个易于完成的子任务。

5.3 多机协作的开放实验平台

目前，硬件平台的集成和智能化系统的相关技术已经可以初步实现个体实验平台的无人化和自主化。未来的发展方向是多机协作开放实验平台的构建，为此需要重点关注4个问题。

（1）多机并行。当前的自动化实验方法大多使用单个机械臂或单体移动机器人，缺乏长序列任务的调度能力，实验吞吐量较低。多机平台采用并行化工作流程和多机协作调度，引入标准化的批量实验操作、数据处理流程、灾备冗余预案，提高工作效率，降低实验的不确定性，实现实验的可控性、资源配置的最优化及灵活性。

（2）开放协作。当无人实验平台无法自主完成复杂的未见的实验任务时，研究人员可通过手动引导的方式跟机器人协作完成宏动—微动实验操作，形成“开放协作、人在环路”的人—机融合实验模式。

（3）实验结果评价。依据特定科学领域，设计合理的无人实验成效评价机制，主要包括：① 科学实验本身的成功率，即与人工实验结果的一致性；② 实验操作是否达到应有的标准（如操作和定位精度）；③ 并行协作平台的实验过程的通量水平和工作效率。

（4）安全保障。大规模开放无人实验平台的安全性可参考无人驾驶技术，从完全封闭环境下的无人参与实验环境，到实验机器人可在研究人员参与的实验环境中自主安全移动，最终在高度自治的基础上实现人—机协作共融。

6 高效、可信的大规模科研协作

在AI驱动科学研究的新范式下，大规模跨领域的协作已成为科学研究的必然模式。智能化科学设施支持数据共享，以便于AI模型的开发与测试，但是需要设计一系列措施对数据开发者的知识产权与利益进行确认与保护。近年来，去中心化科学（DeSci）也备受科研人员关注，DeSci旨在使用Web3工具，包括智能合约和区块链来解决科学研究中的知识产权问题，促进科学数据的共享流通。在智能化科学设施的架构中，区块链技术为打造安全、可信的协作环境提供基础，联邦学习技术能够在去中心化的协作环境中解决数据孤岛问题并保证数据安全和科研效率，互联网群体智能可以将不同的科研模块整合在统一的平台中，实现高效的大规模科研协作。



图8 高效、可信的大规模科研协作系统
Figure 8 Efficient, trustworthy, and large-scale research collaboration system

（1）基于区块链的可信计算。基于区块链的可信计算可对科研协作过程中各方的工作和贡献进行溯源，实现科研协作多方确权。区块链和AI的有效结合将是构建可信的科研协作环境的有效方法。AI技术可以实现区块链的可预测性，并有效检测区块链及其智能合约的漏洞，提升其安全性。反之，区块链技术有望解决AI模型的分布式数据共享和分布式训练推理等问题。

（2）基于联邦学习的科学计算。联邦学习有望实现数据去中心化的科学计算，通过相容激励的机制设计，实现多用户端协作的机器学习模型训练和推理。基于联邦学习的科学计算只要求用户将数据存储在本地，不需要每个用户上传数据，通过迭代的模型聚合实现学习目标，满足大规模科研协作的科学数据及模型的去中心化需求。

（3）网络空间中的群体智能。综合运用数联网、群体决策、语言大模型等技术，科研群体有望在网络空间形成大规模的科研协作，突破传统科研协作的时空局限性，为科研人员推荐相关的研究团队、连接不同领域的科研人员、增强跨学科科研协作、实现对传统科学设施的AI赋能。其关键技术可以概括为：① 基于数字对象的数联网。数联网是基于互联网的虚拟数据网络，通过数字对象有效连接分布式的数据平台，支撑多元异构科学数据的互通互联。以数字对象为核心的数联网为构建AI for Science科研协作平台提供了有效载体[14]。② 群体智能决策策略。网络空间中的人类与机器之间共同协作的群体智能通过“探索—融合—反馈”机制[15]，有望促进开源社区等协作平台的决策效率。③ 基于语言大模型的智能体。ChatGPT等语言大模型为机器与机器之间群体智能科研协作提供了新的契机。由于语言大模型拥有知识、推理能力、思维链功能，使得其可以作为智能体，实现复杂协作任务中的自主管理和调度。

7 总结与展望

AI for Science正在成为驱动科学研究的新范式，引起各国政府、高校和科研机构的高度重视。本文提出了建设智能化科学设施的构想，阐述了“算力支撑—科学引擎—无人实验”3层基础架构和“科学大模型、生成式模拟与反演、高通量自主无人实验、大规模可信协作”4个创新功能，形成高度数字化和智能化的新型智能化科学设施，同时对已有的重大科学设施进行AI赋能。

智能化科学设施的建设不仅有助于解决复杂科学问题、促进跨学科合作创新、开辟新的科学疆域，而且有望对工程技术、未来产业起到重要推动作用。在工程技术方面，智能化科学设施可以提高大规模和复杂工程问题的仿真和推理能力，对复杂工程场景做出更加准确的预测，提高重大工程设备的可靠性和运行效率。在产业方面，智能化科学设施的建设将促进我国产业界承接基础研究的新成果，并充分利用AI和区块链等技术，探索低成本、高可信、标准化的CRO（合同研究组织）科创合作模式及其“风险共担、收益共享”的激励机制，提升重大科技成果的转化效率和质量，对未来产业发展起到支撑作用。

致谢 衷心感谢丁奎岭院士、程津培院士对本文的指导。

参考文献

1 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583-589.

2 Degrave J, Felici F, Buchli J, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. Nature, 2022, 602: 414-419.

3 Zhou Y D, Wang F, Tang J, et al. Artificial intelligence in COVID-19 drug repurposing. The Lancet Digital Health, 2020, 2(12): e667-e676.

4 Coley C W, Thomas III D A, Lummiss J A, et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. Science, 2019, 365: eaax1566.

5 Zhao H T, Chen W, Huang H, et al. A robotic platform for the synthesis of colloidal nanocrystals. Nature Synthesis, 2023, 2 (6): 505-514.

6 Zhang L F, Han J Q, Wang H, et al. Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. Physical Review Letters, 2018, 120(14): 143001.

7 Yao Z P, Sánchez-Lengeling B, Bobbitt N S, et al. Inverse design of nanoporous crystalline reticular materials with deep generative models. Nature Machine Intelligence, 2021, 3(1): 76-86.

8 Lee J, Yoon W, Kim S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 2020, 36(4): 1234-1240.

9 Li Z Y, Kovachki N, Azizzadenesheli K, et al. Fourier neural operator for parametric partial differential equations// International Conference on Learning Representations. Vienna: International Conference on Learning Representations, 2021.

10 Raissi M, Yazdani A, Karniadakis G E. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. Science, 2020, 367: 1026-1030.

11 Cranmer M, Sanchez-Gonzalez A, Battaglia P, et al. Discovering symbolic models from deep learning with inductive biases. Advances in Neural Information Processing Systems, 2020, 33: 17429-17442.

12 Guan S Y, Deng H Y, Wang Y B, et al. NeuroFluid: Fluid dynamics grounding with particle-driven neural radiance fields// International Conference on Machine Learning. Baltimore: ACM, 2022: 7919-7929.

13 Ibarz J, Tan J E, Finn C, et al. How to train your robot with deep reinforcement learning: Lessons we have learned. The International Journal of Robotics Research, 2021, 40(4-5): 698-721.

14 罗超然, 马郓, 景翔, 等. 数据空间基础设施的技术挑战及数联网解决方案. 大数据, 2023, 9(2): 110-121.

Luo C R, Ma Y, Jing X, et al. Internet of data: A solution for dataspace infrastructure and its technical challenges. Big Data Research, 2023, 9(2): 110-121. (in Chinese)

15 Zhang W, Mei H. A constructive model for collective intelligence. National Science Review, 2020, 7(8): 1273-1277.

AI for Science: AI enabled scientific facility transforms fundamental research

YANG Xiaokang[1] XU Yanyan[1] CHEN Lu[1,2] WANG Yunbo[1] GAO Yue[1] TIAN Jidong[1]
YU Kai[1,2] JIN Yaohui[1] MEI Hong[3*]
（1 Artificial Intelligence Institute, Shanghai Jiao Tong University, Shanghai 200240, China;
2 Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China;
3 School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China）

Abstract In recent years, artificial intelligence (AI) has achieved numerous disruptive breakthroughs in frontier scientific and technological fields, such as AlphaFold2 for protein structure prediction, intelligent control of nuclear fusion, and drug design for COVID-19. These achievements indicate that AI for Science is becoming a new paradigm in research. To achieve fundamental scientific innovation and major technological breakthroughs in the era of intelligence, two core issues should be addressed: 1) how to harness the generality and creativity of the new-generation of AI, especially generative AI and large language models (LLMs), to promote the formation of new paradigms; 2) how to empower and transform traditional scientific facilities using AI. To tackle these challenges, this study proposes a concept of AI-enabled scientific facility (AISF) that caters to the requirements of both establishment of totally new intelligent scientific facility and AI empowerment of existing scientific facilities. It aims to construct an infrastructure system for AI for Science, enabling innovative functionalities such as scientific large language models (LLMs), generative simulation and inversion, autonomous intelligent unmanned experiments, and large-scale trustworthy scientific collaboration. These advancements will accelerate scientific discoveries, synthesis of transformative materials, and application of related engineering technologies.

Keywords AI for Science, generative AI, large language models, scientific facility

杨小康 上海交通大学人工智能研究院教授、常务副院长，人工智能教育部重点实验室主任，上海市图像图形学会理事长。主要研究方向为图像处理、机器学习、科学智能。E-mail: xkyang@sjtu.edu.cn

YANG Xiaokang Professor and Executive Deputy Dean of Artificial Intelligence Institute at Shanghai Jiao Tong University, Director of Key Laboratory of Artificial Intelligence of Ministry of Education (MOE) at SJTU, President of Shanghai Image and Graphics Association. His research focuses on image processing, machine learning, and AI for Science. E-mail: xkyang@sjtu.edu.cn

梅 宏 中国科学院院士，发展中国家科学院院士，欧洲科学院外籍院士。高可信软件技术教育部重点实验室（北京大学）主任、上海交通大学人工智能研究院院长。北京大数据先进技术研究院理事长，中国计算机学会理事长。主要从事软件工程和系统软件领域的研究。E-mail: meih@pku.edu.cn

MEI Hong Academician of Chinese Academy of Sciences (CAS), Fellow of the World Academy of Sciences for the advancement of science in developing countries (TWAS), Foreign Member of Academia Europaea, Professor and Director of Key Lab of High Confidence Software Technologies (Peking University) of Ministry of Education (MOE), Director of Artificial Intelligence Institute at Shanghai Jiao Tong University, President of Advanced Institute of Big Data, Beijing, President of China Computer Federation (CCF). His current research covers software engineering and system software. E-mail: meih@pku.edu.cn

■责任编辑：张帆

*Corresponding author